Neurocomputing 168 (2015) 47–54

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Global mutual information-based feature selection approach using single-objective and multi-objective optimization

Min Han *, Weijie Ren

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, Liaoning 116024, China

ARTICLE INFO

Article history:
Received 12 February 2015
Received in revised form
25 April 2015
Accepted 10 June 2015
Communicated by H. Zhang
Available online 18 June 2015

Keywords:
Feature selection
Mutual information
Search strategy
Multi-objective optimization
Time series

ABSTRACT

Feature selection is an important preprocessing step in data mining. Mutual information-based feature selection is a kind of popular and effective approaches. In general, most existing mutual information-based techniques are greedy methods, which are proven to be efficient but suboptimal. In this paper, mutual information-based feature selection is transformed into a global optimization problem, which provides a new idea for solving feature selection problems. First, a single-objective feature selection algorithm combining relevance and redundancy is presented, which has well global searching ability and high computational efficiency. Furthermore, to improve the performance of feature selection, we propose a multi-objective feature selection algorithm. The method can meet different requirements and achieve a tradeoff among multiple conflicting objectives. On this basis, a hybrid feature selection framework is adopted for obtaining a final solution. We compare the performance of our algorithm with related methods on both synthetic and real datasets. Simulation results show the effectiveness and practicality of the proposed method.

1. Introduction

With the development of data acquisition and storage technology, high-dimensional data widely exist in nature [1], finance [2], industry [3–5], biomedicine [6–8] and many other fields, which contain complicated nonlinear relationship among multiple features. Finding potential useful information and building prediction model from high-dimensional data have become one of the most important aspects of data mining and knowledge discovery [9]. Although the high-dimensional data could provide abundant information, it is more and more difficult to establish accurate prediction model along with the increasing dimensionality and scale of dataset [10]. At the same time, the existence of irrelevant and redundant features can easily cover up the effect of important features, which will have negative impact on modeling. Therefore, to deal with these problems, dimensionality reduction approaches, including feature extraction and feature selection, have attracted much attention.

Feature selection [11] is the process of selecting one informative feature subset from the original dataset. It is a key problem in the field of pattern recognition and widely used for processing high-dimensional dataset. Compared to feature extraction, feature selection could maintain physical properties of the original features and has better interpretability, which has been widely used in time series analysis [12] and pattern classification [13]. For example, in order to improve the prediction accuracy and reduce training time, selecting proper training set instance is an important preprocessing step for long-term time series prediction. For the problem of hyperspectral image classification, there are numerous features that can be extracted from the original image. Choosing the most important feature subset not only simplifies the classifier, but also improves the classification accuracy and generalization performance. Thus, feature selection is an essential part for solving real world applications. Currently, researches of feature selection method mainly focus on the following two aspects [14]: defining an appropriate evaluation criterion and designing an effective search strategy.

To realize dimensionality reduction of feature space, correlation analysis measures [15] are necessary. At present, the commonly used methods include Granger causality analysis, Pearson correlation coefficient, canonical correlation analysis (CCA), mutual information (MI), etc. Among them, MI [16] can measure both the linear and nonlinear dependency between features without any prior knowledge, which has been widely applied in feature selection problem. The basic idea of MI-based feature selection algorithm [17] is to select one optimal subset from the original dataset by maximizing the joint MI between the input features and target output. However, when dealing with high-dimensional data, estimation of high-dimensional MI is very difficult and has high computational complexity, which limits the applications of this method. To avoid estimating the joint MI, most of the existing methods adopt low-dimensional MI approximation, such as mutual information feature selection (MIFS) [18],

* Corresponding author.
E-mail address: minhan@dlut.edu.cn (M. Han).

http://dx.doi.org/10.1016/j.neucom.2015.06.016
0925-2312/

minimal redundancy maximal relevance (mRMR) [19], normalized mutual information feature selection (NMIFS) [20], etc. This kind of method has low computational complexity and archives good feature selection results.

Search strategy is another key problem of feature selection [21], which contains exhaustive search, heuristic search, random search, etc. The exhaustive search, that evaluates all possible subsets, is a global optimal search strategy. However, as the dimensionality of input features increases, it becomes impractical and infeasible, because of its high computational complexity. Hence, a variety of heuristic methods with relative low computational complexity have been adopted for feature selection, such as sequential forward selection and backward elimination [22]. But, these algorithms suffer from nesting effect and easily lead to suboptimal results, as they do not take the interdependence among multiple features into consideration. In order to achieve better results, efficient global search strategy is needed. Evolutionary algorithms [23] with global search ability and high computational efficiency are highly suitable for solving the problem of feature selection.

Traditionally, a single objective function is designed for feature selection problem. However, feature selection is a multi-objective problem essentially [24,25]. For example, Peng et al. [19] proposed two objective functions, which are maximal relevance and minimal redundancy, for MI-based feature selection problem. The weighting method can integrate different objectives into a single function, but the weights among multiple objectives are difficult to automatically identify. In addition, the different objectives are usually conflicted, which should be optimized simultaneously to achieve a trade-off between them. Therefore, treating feature selection as a multi-objective problem, that has no additional parameters to be optimized, is more appropriate to meet different goals [26,27].

In this paper, we investigate several commonly used MI-based feature selection algorithms and propose global MI-based feature selection methods based on single-objective and multi-objective optimization. First, a global optimization framework for MI-based feature selection is proposed. Then, a hybrid feature selection algorithm is presented, which combines the efficiency of filters and the high accuracy of wrappers. And the time complexities of the proposed methods are detailed analyzed. Experimental results on both synthetic and real datasets show the performance of the proposed method. The rest of this paper is organized as follows. In Section 2, the background of MI will be introduced and MI-based feature selection methods will be discussed and compared in detail. In Section 3, we will give a detailed presentation for the proposed global MI-based feature selection methods. And the experimental results are analyzed in Section 4. Finally, the conclusions are given in Section 5.

2. Review of mutual information-based feature selection

In this section, we briefly review the definition and estimation of mutual information (MI). Then, detailed analyses of the commonly used MI-based feature selection algorithms are provided.

2.1. Mutual information

In the field of probability theory and information theory, MI is a kind of method for correlation analysis, which is used to measure statistical dependency between random variables. For two continuous random variables X and Y with probability density function $p(x)$ and $p(y)$ respectively, the MI is defined as follows:

$$I(X;Y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{1}$$

where $p(x,y)$ is the joint probability density function of X and Y. In general, the stronger correlation between two random variables is, the larger MI they will have. In a particular case, when the two random variables X and Y are independent, that is $p(x,y)=p(x)p(y)$, the MI will be zero.

Then, extend the pairwise MI to high-dimensional case. Considering the dependency between multiple random variables $\{X_1,X_2,\ldots,X_m\}$ and Y, we can obtain the joint MI as follows:

$$I(X_1,X_2,\ldots,X_m;Y)=\iint\cdots\int p(x_1,x_2,\ldots,x_m,y)\log\frac{p(x_1,x_2,\ldots,x_m,y)}{p(x_1,x_2,\ldots,x_m)p(y)}dx_1dx_2\cdots dx_mdy \tag{2}$$

where $p(x_1,x_2,\ldots,x_m)$ and $p(x_1,x_2,\ldots,x_m,y)$ are joint probability density functions. Unlike the pairwise MI, the joint MI not only concerns the dependency between $\{X_1,X_2,\ldots,X_m\}$ and Y, but also involves the internal correlation of $\{X_1,X_2,\ldots,X_m\}$. Therefore, the joint MI is highly suited for solving feature selection problems.

In practical applications, the accuracy of MI estimation directly influences the identification of dependency between random variables. Currently, the commonly used MI estimators include histogram method [28], kernel method [29], k-nearest neighbor method [30], etc. Every MI estimator has its advantages and scope of applications. In this paper, we adopt the MI estimator based on copula entropy [31], which can avoid the estimation of probability density functions. The MI estimation is transformed into the estimation of copula entropy, which can be applied to estimate both pairwise MI and joint MI efficiently.

2.2. Evaluation criteria

The basic idea of MI-based feature selection algorithm is to select the best subset S from the original dataset F by maximizing the joint MI between S and target output Y. This scheme, namely maximal dependency (MD) criterion [21], is shown as follows:

$$MD:\ \max\{I(S;Y)\} \tag{3}$$

where $S\subseteq F$ is the candidate feature subset, and Y is the target output. When applying MD criterion, accurate estimation of joint MI is an essential condition. However, as the dimensionality of input features increases, the accuracy of MI estimator gradually reduces and the computational complexity rapidly increases. Therefore, to avoid estimating joint MI, many MI-based feature selection methods use low-dimensional MI to approximate the MD criterion.

The simplest approach to realize MI-based feature selection is maximal relevance (MR) criterion. The idea of MR criterion is to select the first m features from the original dataset F by maximizing the pairwise MI between X_i and Y. The objective function is defined as follows:

$$MR:\ \max\{I(X_i;Y)\} \tag{4}$$

where $X_i\in F$ is the candidate feature. MR criterion is easy to implement and has high computational efficiency. The major drawback is that the MR criterion does not consider the redundancy between input features.

On the basis of maximal relevance, Battiti [18] introduces the penalty term of redundancy and proposes mutual information feature selection (MIFS) criterion, which is shown as follows:

$$MIFS:\ \max\left\{I(X_i;Y)-\beta\sum_{X_j\in S}I(X_i;X_j)\right\} \tag{5}$$

where $X_i\in F\backslash S$ is the candidate feature, and $X_j\in S$ is the selected feature. β is a balance parameter, that controls the magnitude of the redundancy penalty term. This criterion considers both relevance and redundancy between features at the same time, but it is heavily influenced by the parameter β.